各期間毎に目標をもつ多段決定過程 (不確実性下における意思決定問題)

AUTHOR(S):

吉良, 知文; 藤田, 敏治

CITATION:

吉良, 知文 ...[et al]. 各期間毎に目標をもつ多段決定過程 (不確実性下における意思決定問題). 数理解析研究所講究録 2011, 1734: 228-235

ISSUE DATE:

2011-03

URL:

http://hdl.handle.net/2433/170756

数理解析研究所講究録
第1734巻 2011年 228-235

各期間毎に目標をもつ多段決定過程

九州大学大学院数理学府　吉良 知文 (Akifumi Kira)
Graduate School of Mathematics, Kyushu University
九州工業大学大学院工学研究院　藤田 敏治 (Toshiharu Fujita)
Graduate School of Engineering, Kyushu Institute of Technology

1 はじめに

マルコフ決定過程において, 期待値だけでなく確率基準の評価を考えるのは自然である. 現実に人間は期待値を最大化するよりも, 望ましい状況, あるいは最低限の要求が満たされるように行動することが多々ある. このような意思決定の原理は "satisfying approach" と呼ばれている [11]. 例えば, ポートフォリオの運用者はしばしば期待値よりも確率に興味がある. このような観点から, 各段で得られる利得の (割引) 総和が目標値以上 (以下) になる確率を最大 (最小) 化する**閾値確率問題**が多くの研究者によって研究されており, "target-level criterion" あるいは "minimizing risk model" とも呼ばれている [4, 9, 10, 12, 13, 14, 15, 16]. 特に, Wu and Lin [16] は先行論文の不備を指摘した上で可算の状態空間と利得集合をもつマルコフ決定過程として定式化し, 有限期間および無限期間の問題に対する最適値関数が閾値の分布関数であることを示し, 有限期間の問題に対して拡大状態空間上でマルコフ性をもつ確定的な政策の中に最適政策が存在することを示した. 無限期間の問題に対しては, Ohtsubo and Toyonaga [10] によって右連続な最適定常政策の存在のための2つの十分条件が与えられた. Boda and Filar [3] は有限期間の閾値確率問題を応用して, 動的ポートフォリオ配分問題といった多期間の問題に適した動的リスク測度について議論し, Value-at-Risk (VaR) および Conditional Value-at-Risk (CVaR) の multi-stage version を提唱している.

有限期間の問題に的を絞ると上述の閾値確率問題は, 終端時刻 N において目標を達成できるか否かが問題であり, **終端型**の評価である. 本稿ではこの拡張として, 各期間毎にそれまでに得られる利得の集積値に課せられる目標値が設定されている状況を想定する. 全ての目標を達成する確率を最大化する**非終端型**の閾値確率問題と目標達成回数の期待値を最大化する2つの問題を導入し, 動的計画法による再帰式を導く. 特に前者は**流動性リスク**を評価に加えた拡張と解釈することができ, 「黒字倒産」などという言葉が話題となった近年の経済状況を考慮しても必要かつ自然な拡張であると思われる.

2 有限段マルコフ決定過程

本稿では数値計算のために有限性を意識したマルコフ決定過程 $\mathcal{D}$ を扱う.

$$\mathcal{D} = \left(X, (U, \{U_n(\cdot)\}_0^{N-1}), (\{r_n\}_0^{N-1}, r_G), p \right)$$

1. $N\ (\geq 2)$ は**段 (期) の総数**.

2. $X = \{s_1, s_2, \ldots, s_m\}$ は有限**状態空間**. 時刻 n に確率的に生じる状態を $X_n\,(\in X)$ で表し, 実現した状態を x_n で表す $(n = 0, 1, \ldots, N)$.
3. $U = \{a_1, a_2, \ldots, a_k\}$ は有限**決定空間**. $u_n\,(\in U)$ は時刻 n での決定を表す $(n = 0, 1, \ldots, N-1)$. $U_n : X \to 2^U \setminus \{\phi\}$ は点対集合値で, $U_n(x)$ は時刻 n での状態が x であるときに実行可能な決定全体を表す. また, $G_r(U_n)$ を $U_n(\cdot)$ のグラフとする :

$$G_r(U_n) = \{(x, u) \,|\, u \in U_n(x),\ x \in X\}$$

4. $r_n : G_r(U_n) \to D \subset \mathbb{R}$ $(n = 0, 1, \ldots, N-1)$ は第 n **利得関数**. 時刻 n に状態 x_n において決定 $u_n\,(\in U_n(x_n))$ を選ぶと利得 $r_n(x_n, u_n)$ を得る. $r_G : X \to D$ は**終端利得関数**. 最終時刻 N では状態 x_N で利得 $r_G(x_N)$ を得る.
5. $p = \{p(\cdot\,|x, u)\}$ は定常な**マルコフ推移法則**. $p(y|x_n, u_n)$ は状態 x_n において決定 u_n を選んだときに, 次状態 X_{n+1} が $y\,(\in X)$ になる条件付き確率である. この確率的推移を $X_{n+1} \sim p(\cdot\,|x_n, u_n)$ と表現する.

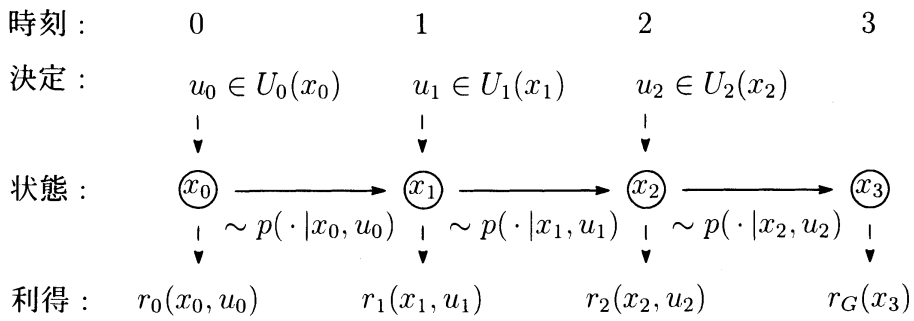

図 1: 有限段マルコフ決定過程 $(N = 3)$

3 原問題

各段 (期) で達成すべき目標として所与の区間 $I_n \subset \mathbb{R}$ $(n = 0, 1, \ldots. N)$ が与えられている. 時刻 0 から時刻 n までに得られる利得の集積値がこの区間 I_n に収まることが望ましい. ここで利得の集積値として結合型の評価値 (Iwamoto [7]) を考える. この集積値を簡単に $\bigcirc_{t=0}^{n} r_t$ と書くことにし, 次のように定義する.

$$\bigcirc_{t=0}^{n} r_t := \begin{cases} r_0(X_0, U_0) & \text{if } n = 0 \\ r_0(X_0, U_0) \circ r_1(X_1, U_1) & \text{if } n = 1 \\ \quad\vdots & \quad\vdots \\ r_0(X_0, U_0) \circ r_1(X_1, U_1) \circ \cdots \circ r_{N-1}(X_{N-1}, U_{N-1}) & \text{if } n = N-1 \\ r_0(X_0, U_0) \circ r_1(X_1, U_1) \circ \cdots \circ r_{N-1}(X_{N-1}, U_{N-1}) \circ r_G(X_N) & \text{if } n = N\,. \end{cases}$$

ただし, $\circ$ は利得関数の値域 D 上で定義された結合律を満たす二項演算子で, 左単位元 $e \in D$ をもつものとする. 時刻 0 に始まる N 期間の**一般政策**全体を $\Sigma^{(0,N)}$ で表す :

$$\Sigma^{(0,N)} := \left\{ \sigma = (\sigma_0, \ldots, \sigma_{N-1}) \;\middle|\; \begin{array}{l} \sigma_n : X^{n+1} \to U, \\ \sigma_n(x_0, \ldots, x_n) \in U_n(x_n),\ \forall (x_0, \ldots, x_n) \in X^{n+1}, \\ n = 0, 1, \ldots, N-1 \end{array} \right\}.$$

一般政策 $\sigma = (\sigma_0, \sigma_1, \ldots, \sigma_{N-1}) \in \Sigma^{(0,N)}$ を採用した意思決定者は, 時刻 n において, それまでの状態列 $(x_0, x_1, \ldots, x_n)$ に依存した決定 $u_n = \sigma_n(x_0, x_1, \ldots, x_n)$ を選択する. このとき, 任意の $x \in X$ に対して, x が初期状態として与えられたときに全ての目標を達成できる確率を最大化する非終端型閾値確率問題 $\mathrm{P}(x)$ を考える.

$$\mathrm{P}(x)\quad \begin{array}{ll} \text{Maximize} & \mathrm{P}^{\sigma}\Big(\bigcirc_{t=0}^{n} r_t \in I_n,\ n = 0, 1, \ldots, N \,\Big|\, X_0 = x \Big) \\ \text{subject to} & \text{(i) } X_{n+1} \sim p(\cdot\,|x_n, u_n), \quad n = 0, 1, \ldots, N-1 \\ & \text{(ii) } \sigma \in \Sigma^{(0,N)}. \end{array}$$

ただし, P^{σ} は一般政策 σ が採用された下での条件付確率を表す. また, 異なる評価基準として, 目標達成回数の期待値を最大化する問題 $\mathrm{Q}(x)$ が考えられる.

$$\mathrm{Q}(x)\quad \begin{array}{ll} \text{Maximize} & \mathrm{E}^{\sigma}\Big[\#\Big\{ n \in \mathcal{N} \,\Big|\, \bigcirc_{t=0}^{n} r_t \in I_n \Big\} \,\Big|\, X_0 = x \Big] \\ \text{subject to} & \text{(i) } X_{n+1} \sim p(\cdot\,|x_n, u_n), \quad n = 0, 1, \ldots, N-1 \\ & \text{(ii) } \sigma \in \Sigma^{(0,N)}. \end{array}$$

E^{σ} は一般政策 σ が採用された下での条件付期待値を表し, $\mathcal{N} = \{0, 1, \ldots, N\}$ とする. 通常の閾値確率問題においては未来閾値に基づく埋め込みによって再帰式が得られるが, 目標が複数ある問題に対しては過去集積値に基づく埋め込み [7, 13] が有効である.

4 埋め込み問題と再帰式

両問題の統一的表現のために二項演算子 $\times$, $+$ のどちらか一方を表す $\bullet$ と, 定義関数 :

$$\psi_n(z) := \mathbf{1}_{I_n}(z) = \begin{cases} 1 & \text{if } z \in I_n \\ 0 & \text{otherwise} \end{cases} \qquad z \in D$$

を導入する. このとき, 両問題の目的関数はそれぞれ次の期待値で表される.

$$\mathrm{E}^{\sigma}\Big[\psi_0\Big(\bigcirc_{t=0}^{0} r_t \Big) \bullet \psi_1\Big(\bigcirc_{t=0}^{1} r_t \Big) \bullet \cdots \bullet \psi_N\Big(\bigcirc_{t=0}^{N} r_t \Big) \,\Big|\, X_0 = x \Big] \tag{1}$$

$\bullet$ が $\times$ であるとき $\mathrm{P}(x)$, $+$ であるとき $\mathrm{Q}(x)$ の目的関数である. また, Λ_0 を $(D, \circ)$ の左単位元 e を含む有限集合とし, パラメータ $\lambda \in \Lambda_0$ が埋め込まれた次の期待値を考える.

$$\mathrm{E}^{\sigma}\Big[\psi_0\Big(\lambda \circ \bigcirc_{t=0}^{0} r_t \Big) \bullet \psi_1\Big(\lambda \circ \bigcirc_{t=0}^{1} r_t \Big) \bullet \cdots \bullet \psi_N\Big(\lambda \circ \bigcirc_{t=0}^{N} r_t \Big) \,\Big|\, X_0 = x \Big] \tag{2}$$

$\lambda = e$ のとき (2) 式は (1) 式に等しくなるので, 目的関数を (2) 式に置き換えたパラメトリックな問題は**埋め込み問題**と呼ばれる. さらに結合律を満たす仮定から,

$$\lambda \circ \bigcirc_{t=0}^{k} r_t = \lambda \circ \bigcirc_{t=0}^{n-1} r_t \circ \bigcirc_{t=n}^{k} r_t, \quad 0 < n \le k$$

という表現ができる. ここで導入した記号 $\bigcirc_{t=n}^{k} r_t$ は時刻 n から時刻 k までに得られる利得の集積値を表す. また, $\lambda \circ \bigcirc_{t=0}^{n-1} r_t$ が取り得る値の集合を Λ_n とする $(n = 1, 2, \ldots, N)$.

$$\Lambda_n := \big\{ \lambda \circ r_0(x_0, u_0) \circ \cdots \circ r_{n-1}(x_{n-1}, u_{n-1}) \,\big|\, \lambda \in \Lambda_0,\ (x_0, \ldots, x_{n-1}) \in X^n,\ \sigma \in \Sigma^{(0,n)} \big\}.$$

任意の $\lambda \in \Lambda_n$ と任意の状態 x に対して, 時刻 n に状態 x から始まる $(N-n)$ 期間の部分問題 $\mathrm{R}_n(x\,;\lambda)$ を次のように定義し, その最適値を $V_n(x\,;\lambda)$ と表す $(n = 0, 1, \ldots, N-1)$.

$$\mathrm{R}_n(x\,;\lambda)\quad\begin{array}{ll}\text{Maximize} & \mathrm{E}^{\sigma}\Big[\,\psi_n\Big(\lambda\circ\bigcirc_{t=n}^{n}r_t\Big)\bullet\psi_{n+1}\Big(\lambda\circ\bigcirc_{t=n}^{n+1}r_t\Big)\bullet\cdots\bullet\psi_N\Big(\lambda\circ\bigcirc_{t=n}^{N}r_t\Big)\,\Big|\,X_n=x\,\Big]\\ \text{subject to} & (\mathrm{i})_n\ X_{t+1}\sim p(\,\cdot\,|x_t,u_t),\quad t=n,\,n+1,\,\ldots,N-1\\ & (\mathrm{ii})_n\ \sigma\in\Sigma^{(n,N-n)}\end{array}$$

ただし, $\Sigma^{(n,N-n)}$ は時刻 n に始まる $(N-n)$ 期間の一般政策全体である：

$$\Sigma^{(n,N-n)}:=\left\{\sigma=(\sigma_n,\,\ldots,\sigma_{N-1})\;\middle|\;\begin{array}{l}\sigma_t:X^{t+1-n}\to U,\\ \sigma_t(x_n,\,\ldots,x_t)\in U_t(x_t),\ \forall(x_n,\,\ldots,x_t)\in X^{t+1-n},\\ t=n,\,\ldots,N-1\end{array}\right\}.$$

開始時刻 n, 始発状態 x, パラメータ λ を変化させて得られる一連の部分問題群に対する最適値関数の列 $V_n:X\times\Lambda_n\to[0,1]$, $n=0,\,1,\,\ldots,N-1$ の間に再帰式を導くことを考える. また, 終了期 $(n=N)$ に対する値関数 $V_N:X\times\Lambda_N\to[0,1]$ を次のように定める.

$$V_N(x\,;\lambda):=\psi_N(\,\lambda\circ r_G(x)\,).$$

定理 4.1.

$$V_N(x\,;\lambda)=\psi_N(\,\lambda\circ r_G(x)\,),\quad(x,\lambda)\in X\times\Lambda_N.$$

$$V_n(x\,;\lambda)=\max_{u\in U_n(x)}\Big\{\psi_n(\,\lambda\circ r_n(x,u)\,)\bullet\sum_{y\in X}V_{n+1}(y\,;\lambda\circ r_n(x,u))p(y|x,u)\Big\},$$
$$(x,\lambda)\in X\times\Lambda_n,\quad n=0,\,1,\,\ldots,N-1.$$

定理 4.2. $\bar\pi_n^*:X\times\Lambda_n\to U$ $(n=0,\,1,\,\ldots,N-1)$ を次のように定義する.

$$\bar\pi_n^*(x,\lambda)\in\arg\max_{u\in U_n(x)}\Big\{\psi_n(\,\lambda\circ r_n(x,u)\,)\bullet\sum_{y\in X}V_{n+1}(y\,;\lambda\circ r_n(x,u))p(y|x,u)\Big\}.\qquad(3)$$

元 $\lambda_0\in\Lambda_0$ を任意に 1 つ選び, 一般政策 $\sigma^*=(\sigma_0^*,\sigma_1^*,\,\ldots,\sigma_{N-1}^*)$ を次のように定める：

$$\begin{aligned}\sigma_0^*&=\sigma_0^*(x_0):=\bar\pi_0^*(x_0,\lambda_0)\\ \sigma_1^*&=\sigma_1^*(x_0,x_1):=\bar\pi_1^*(x_1,\lambda_0\circ r_0(x_0,\sigma_0^*))\\ \sigma_2^*&=\sigma_2^*(x_0,x_1,x_2):=\bar\pi_2^*(x_2,\lambda_0\circ r_0(x_0,\sigma_0^*)\circ r_1(x_1,\sigma_1^*))\\ &\ \ \vdots\\ \sigma_{N-1}^*&=\sigma_{N-1}^*(x_0,\,\ldots,x_{N-1}):=\bar\pi_{N-1}^*(x_{N-1},\lambda_0\circ r_0(x_0,\sigma_0^*)\circ\cdots\circ r_{N-2}(x_{N-2},\sigma_{N-2}^*)).\end{aligned}\qquad(4)$$

このとき, σ^* は 問題群 $\{\mathrm{R}_0(x\,;\lambda_0)\,|\,x\in X\}$ に対する最適政策である.

再帰式を解くことで得られる値 $V_0(x\,;e)$ は $\bullet$ が $\times$ であるとき問題 $\mathrm{P}(x)$, $+$ であるとき $\mathrm{Q}(x)$ の最適値である. それぞれの最適政策は (4) 式において $\lambda_0=e$ とすることで得られる. ただし, e は $(D,\circ)$ の 1 つの左単位元を表す.

5 証明の概略

拡張されたマルコフ決定過程 $\bar{\mathcal{D}}=\big(S,\,(\mathcal{A},\{\mathcal{A}_n(\cdot)\}_0^{N-1}),\,(\{\bar r_n\}_0^{N-1},\bar r_G),\,q\big)$ を考える：

1. S は $S := X \times \bigcup_{n=0}^{N} \Lambda_n$ で定義され, **拡大状態空間**とよぶ.
2. $\mathcal{A}$ は U と同一の決定空間で, 可能決定空間 $\mathcal{A}_n(\cdot)$ は任意の $s_n = (x_n, \lambda_n) \in S$ に対し第1成分のみに依存して, $\mathcal{A}_n(s_n) := U_n(x_n)$ で与えられる.
3. 拡大状態空間上の利得関数 $\bar{r}_n : G_r(\mathcal{A}_n) \to \{0,1\}$ $(n = 0, 1, \ldots, N-1)$, および終端利得関数 $\bar{r}_G : S \to \{0,1\}$ を次のように定義する.

$$\bar{r}_n(s_n, u_n) := \psi_n(\lambda_n \circ r_n(x_n, u_n)), \quad (s_n, u_n) = (x_n, \lambda_n, u_n) \in G_r(\mathcal{A}_n)$$
$$\bar{r}_G(s_N) := \psi_N(\lambda_N \circ r_G(x_N)), \quad s_N = (x_N, \lambda_N) \in S$$

4. 拡大状態空間上の非定常マルコフ推移法則 $q = \{q_n(\cdot\,|s,u)\}$ を次で定義する.

$$q_{n+1}\big(\underbrace{(x_{n+1}, \lambda_{n+1})}_{=s_{n+1}} \,\big|\, \underbrace{(x_n, \lambda_n)}_{=s_n}, u_n\big) := \begin{cases} p(x_{n+1}|x_n, u_n) & \lambda_{n+1} = \lambda_n \circ r_n(x_n, u_n) \\ 0 & \text{otherwise} \end{cases}$$

この状態推移を $S_{n+1} \sim q_{n+1}(\cdot\,|s_n, u_n)$ と表す.

さて, 任意の $(x, \lambda) \in X \times \Lambda_n \subset S$ に対して, 期待値最大化問題 $\bar{\mathrm{R}}_n(x;\lambda)$ を考えよう.

$$\bar{\mathrm{R}}_n(x;\lambda) \quad \begin{array}{ll} \text{Maximize} & \mathrm{E}^{\sigma}\big[\bar{r}_n(S_n, U_n) \bullet \cdots \bullet \bar{r}_{N-1}(S_{N-1}, U_{N-1}) \bullet \bar{r}_G(S_N) \,|\, S_n = (x, \lambda)\big] \\ \text{subject to} & (\mathrm{i})'_n \ S_{t+1} \sim q_{t+1}(\cdot\,|s_t, u_t), \quad t = n, n+1, \ldots, N-1 \\ & (\mathrm{ii})'_n \ \sigma \in \bar{\Sigma}^{(n,N-n)}. \end{array}$$

ただし, $\bar{\Sigma}^{(n,N-n)}$ は S 上の一般政策全体を表し, $\Sigma^{(n,N-n)}$ を本質的に含んでいる :

$$\bar{\Sigma}^{(n,N-n)} = \left\{ \bar{\sigma} = (\bar{\sigma}_n, \ldots, \bar{\sigma}_{N-1}) \;\middle|\; \begin{array}{l} \bar{\sigma}_t : S^{t+1-n} \to \mathcal{A}_t, \\ \bar{\sigma}_t(s_n, \ldots, s_t) \in \mathcal{A}_t(s_t), \forall (s_n, \ldots, s_t) \in S^{t+1-n}, \\ t = n, n+1, \ldots, N-1 \end{array} \right\}.$$

ここで, 制約条件 $(\mathrm{ii})'_n$ のみを $(\mathrm{ii})_n$ に置き換えると問題 $\bar{\mathrm{R}}_n(x;\lambda)$ は $\mathrm{R}_n(x;\lambda)$ と同値であることが $\bar{\mathcal{D}}$ の定義から直ちに得られる. したがって, $\bar{\mathrm{R}}_n(x;\lambda)$ は $\mathrm{R}_n(x;\lambda)$ の緩和問題となっている. この問題は $\bullet$ が $+$ であるとき, マルコフ決定過程で最もよく知られている**加法型評価**であり, $\times$ であるとき**非負値乗法型評価**[6, 8] である. よって, $\mathrm{R}_n(x;\lambda)$ の最適値を $\bar{V}_n(x;\lambda)$ と表すことにすると次の再帰式が成り立つ.

$$\bar{V}_N(s) = \bar{r}_G(s), \quad s \in X \times \Lambda_N$$
$$\bar{V}_n(s) = \max_{u \in \mathcal{A}_n(s)} \Big\{ \bar{r}_n(s,u) \bullet \sum_{s' \in S} \bar{V}_{n+1}(s') q_{n+1}(s'|s,u) \Big\}, \ s \in X \times \Lambda_n, \ n = 0, 1, \ldots, N-1.$$

これは定理 4.1 の再帰式と同一であることが $\bar{\mathcal{D}}$ の定義から容易に分かる. したがって, (3) 式によって定まる S 上のマルコフ政策 $\bar{\pi}^* = (\bar{\pi}^*_0, \bar{\pi}^*_1, \ldots, \bar{\pi}^*_{N-1})$ は緩和問題に対する最適政策, すなわち全ての問題 $\{\bar{\mathrm{R}}_0(x;\lambda) \,|\, x \in X, \lambda \in \Lambda_0\}$ に対して最適値を与える政策である. ゆえに各 $\lambda_0 \in \Lambda_0$ に対して, 最適政策 $\bar{\pi}^*$ と常に同じ決定を選択するように(4) 式によって定められた一般政策 σ^* は問題群 $\{\bar{\mathrm{R}}_0(x;\lambda_0) \,|\, x \in X\}$ に対して最適値を与えることになる. 各問題 $\bar{\mathrm{R}}_0(x;\lambda)$ の最適化が本来の制約条件を満たす X 上の一般政策によって達成されることになる. このことと段の総数 N の任意性より $\bar{V}_n(x;\lambda) \equiv V_n(x;\lambda)$ となる. 以上により定理 4.1 および定理 4.2 が証明される.

6 例題

$$\text{Maximize} \quad \mathrm{P}^{\sigma}\left(\sum_{t=0}^{n} r_t \in I_n,\ n=0,\ 1,\ 2,\ 3 \,\middle|\, X_0 = x \right)$$

$$\text{subject to} \quad \text{(i) } X_{n+1} \sim p(\,\cdot\,|x_n, u_n), \quad n=0,\ 1,\ 2 \quad \text{(ii) } \sigma \in \Sigma^{(0,3)}.$$

区間列： $I_0 = I_1 = I_2 = I_3 = [0, \infty)$, 状態空間： $X = \{s_1, s_2\}$

決定空間および可能決定空間： $U \equiv U_0(x) \equiv U_1(x) \equiv U_2(x) \equiv \{a_1, a_2\}$

利得関数：

$r_0(x_0, u_0)$

$x_0 \setminus u_0$	a_1	a_2
s_1	1	2
s_2	3	1

$r_1(x_1, u_1)$

$x_1 \setminus u_1$	a_1	a_2
s_1	1	-3
s_2	2	-1

$r_2(x_2, u_2)$

$x_1 \setminus u_1$	a_1	a_2
s_1	-2	-4
s_2	-3	-1

$r_G(x_3)$

x_2	$r_G(x_2)$
s_1	5
s_2	-3

マルコフ推移法則：

$p(y|x, u)$

$y \setminus (x, u)$	(s_1, a_1)	(s_1, a_2)	(s_2, a_1)	(s_2, a_2)
s_1	0.3	0.7	0.6	0.5
s_2	0.7	0.3	0.4	0.5

この問題に対する定理 4.1 は以下のようになる.

$$V_3(x;\lambda) = \mathbf{1}_{[0,\infty)}(\lambda + r_G(x)), \quad (x,\lambda) \in X \times \Lambda_3 \tag{5}$$

$$V_2(x;\lambda) = \max_{u \in \{a_1,a_2\}} \mathbf{1}_{[0,\infty)}(\lambda + r_2(x,u)) \sum_{y \in \{s_1,s_2\}} V_3(y;\lambda + r_2(x,u))p(y|x,u), \quad (x,\lambda) \in X \times \Lambda_2 \tag{6}$$

$$V_1(x;\lambda) = \max_{u \in \{a_1,a_2\}} \mathbf{1}_{[0,\infty)}(\lambda + r_1(x,u)) \sum_{y \in \{s_1,s_2\}} V_2(y;\lambda + r_1(x,u))p(y|x,u), \quad (x,\lambda) \in X \times \Lambda_1 \tag{7}$$

$$V_0(x;\lambda) = \max_{u \in \{a_1,a_2\}} \mathbf{1}_{[0,\infty)}(\lambda + r_0(x,u)) \sum_{y \in \{s_1,s_2\}} V_1(y;\lambda + r_0(x,u))p(y|x,u), \quad (x,\lambda) \in X \times \Lambda_0. \tag{8}$$

$\Lambda_0 = \{0\}$ とすると, 過去値集合列 $\{\Lambda_n\}$ は次のように定まる.

$$\Lambda_1 = \{\, \lambda + r_0(x_0,u_0) \,|\, \lambda \in \Lambda_0,\ x_0 \in X,\ u_0 \in U \,\} = \{1,\ 2,\ 3\}$$
$$\Lambda_2 = \{\, \lambda + r_1(x_1,u_1) \,|\, \lambda \in \Lambda_1,\ x_1 \in X,\ u_1 \in U \,\} = \{-2,-1,\ 0,\ 1,\ 2,\ 3,\ 4,\ 5\}$$
$$\Lambda_3 = \{\, \lambda + r_2(x_2,u_2) \,|\, \lambda \in \Lambda_2,\ x_2 \in X,\ u_2 \in U \,\} = \{-6,-5,-4,-3,-2,-1,\ 0,\ 1,\ 2,\ 3,\ 4\}.$$

(5) 式より,

$$V_3(s_1;\lambda) = \mathbf{1}_{[0,\infty)}(\lambda+5) = \begin{cases} 1 & \text{if } \lambda \in \{-5,-4,-3,-2,-1,\ 0,\ 1,\ 2,\ 3,\ 4\} \\ 0 & \text{if } \lambda = -6 \end{cases}$$

$$V_3(s_2;\lambda) = \mathbf{1}_{[0,\infty)}(\lambda-3) = \begin{cases} 1 & \text{if } \lambda \in \{3,\ 4\} \\ 0 & \text{if } \lambda \in \{-6,-5,-4,-3,-2,-1,\ 0,\ 1,\ 2\}. \end{cases}$$

次に (6) 式を用いて $V_2(x,\lambda)$ を評価する. 例えば, $V_2(s_1;4)$ の値は次のようにして得られる.

$$\begin{aligned} V_2(s_1;4) &= \max_{u \in \{a_1,a_2\}} \mathbf{1}_{[0,\infty)}(4 + r_2(s_1,u)) \sum_{y \in \{s_1,s_2\}} V_3(y;4 + r_2(s_1,u))p(y|s_1,u) \\ &= \left\{ \mathbf{1}_{[0,\infty)}(4 + r_2(s_1,a_1)) \sum_{y \in \{s_1,s_2\}} V_3(y;4 + r_2(s_1,a_1))p(y|s_1,a_1) \right\} \\ &\qquad \vee \left\{ \mathbf{1}_{[0,\infty)}(4 + r_2(s_1,a_2)) \sum_{y \in \{s_1,s_2\}} V_3(y;4 + r_2(s_1,a_2))p(y|s_1,a_2) \right\} \end{aligned}$$

$$= \left\{ \mathbf{1}_{[0,\infty)}(2) \times (V_3(s_1;2) \times p(s_1|s_1,a_1) + V_3(s_2;2) \times p(s_2|s_1,a_1) \right\}$$
$$\vee \left\{ \mathbf{1}_{[0,\infty)}(0) \times (V_3(s_1;1) \times p(s_1|s_1,a_2) + V_3(s_2;0) \times p(s_2|s_1,a_2) \right\}$$
$$= \left\{ 1 \times (1 \times 0.3 + 0 \times 0.7) \right\} \vee \left\{ 1 \times (1 \times 0.7 + 0 \times 0.3) \right\} = 0.3 \vee 0.7$$
$$= 0.7, \qquad \bar{\pi}_2^*(s_1;4) = a_2 .$$

同様の計算を行うことで全ての $(x,\lambda) \in X \times \Lambda_2$ に対して $V_2(x;\lambda)$ が求まる:

$$V_2(s_1;\lambda) = \begin{cases} 0.0 & \text{if } \lambda \in \{-2,-1,0,1\} \\ 0.3 & \text{if } \lambda \in \{2,3\} \\ 0.7 & \text{if } \lambda = 4 \\ 1.0 & \text{if } \lambda = 5, \end{cases} \qquad \bar{\pi}_2^*(s_1;\lambda) = \begin{cases} \text{either} & \text{if } \lambda \in \{-2,-1,0,1\} \\ a_1 & \text{if } \lambda \in \{2,3,5\} \\ a_2 & \text{if } \lambda = 4 \end{cases} \tag{9}$$

$$V_2(s_2;\lambda) = \begin{cases} 0.0 & \text{if } \lambda = \{-2,-1,0\} \\ 0.5 & \text{if } \lambda = \{1,2\} \\ 0.6 & \text{if } \lambda = 3 \\ 1.0 & \text{if } \lambda \in \{4,5\}, \end{cases} \qquad \bar{\pi}_2^*(s_2;\lambda) = \begin{cases} \text{either} & \text{if } \lambda = \{-2,-1,0\} \\ a_1 & \text{if } \lambda = 3 \\ a_2 & \text{if } \lambda = \{1,2,4,5\}. \end{cases} \tag{10}$$

同様に, (7) 式, (9) 式, および (10) 式から, $V_1(x;\lambda)$ を求めると,

$$V_1(s_1;\lambda) = \begin{cases} 0.44 & \text{if } \lambda = 1 \\ 0.51 & \text{if } \lambda = 2 \\ 0.91 & \text{if } \lambda = 3, \end{cases} \qquad \bar{\pi}_1^*(s_1;\lambda) = a_1,\ \lambda \in \{1,2,3\} \tag{11}$$

$$V_1(s_2;\lambda) = \begin{cases} 0.42 & \text{if } \lambda = 1 \\ 0.82 & \text{if } \lambda = 2 \\ 1.00 & \text{if } \lambda = 3, \end{cases} \qquad \bar{\pi}_1^*(s_2;\lambda) = a_1,\ \lambda \in \{1,2,3\} \tag{12}$$

が得られる. (8) 式, (11) 式, および (12) 式から求める最適値関数が得られる.

$$V_0(s_1;0) = 0.603, \quad \bar{\pi}_0^*(s_1;0) = a_2, \qquad V_0(s_2;0) = 0.946, \quad \bar{\pi}_0^*(s_2;0) = a_1 .$$

最後に定理 4.2 より, 上で求めた $\bar{\pi}^* = (\bar{\pi}_0^*, \bar{\pi}_1^*, \bar{\pi}_2^*)$ から最適な一般政策 $\sigma^* = (\sigma_0^*, \sigma_1^*, \sigma_2^*)$ が次のように構成される.

$$\sigma_0^*(s_1) = \bar{\pi}_0^*(s_1;0) = a_2$$
$$\sigma_0^*(s_2) = \bar{\pi}_0^*(s_2;0) = a_1$$
$$\sigma_1^*(s_1,s_1) = \bar{\pi}_1^*(s_1;r_0(s_1,\sigma_0^*(s_1))) = \bar{\pi}_1^*(s_1;2) = a_1$$
$$\sigma_1^*(s_2,s_1) = \bar{\pi}_1^*(s_1;r_0(s_2,\sigma_0^*(s_2))) = \bar{\pi}_1^*(s_1;3) = a_1$$
$$\sigma_1^*(s_1,s_2) = \bar{\pi}_1^*(s_2;r_0(s_1,\sigma_0^*(s_1))) = \bar{\pi}_1^*(s_2;2) = a_1$$
$$\sigma_1^*(s_2,s_2) = \bar{\pi}_1^*(s_2;r_0(s_2,\sigma_0^*(s_2))) = \bar{\pi}_1^*(s_2;3) = a_1$$
$$\sigma_2^*(s_1,s_1,s_1) = \bar{\pi}_2^*(s_1;r_0(s_1,\sigma_0^*(s_1))+r_1(s_1,\sigma_1^*(s_1,s_1))) = \bar{\pi}_2^*(s_1;3) = a_1$$
$$\sigma_2^*(s_2,s_1,s_1) = \bar{\pi}_2^*(s_1;r_0(s_2,\sigma_0^*(s_2))+r_1(s_1,\sigma_1^*(s_2,s_1))) = \bar{\pi}_2^*(s_1;4) = a_2$$
$$\sigma_2^*(s_1,s_2,s_1) = \bar{\pi}_2^*(s_1;r_0(s_1,\sigma_0^*(s_1))+r_1(s_2,\sigma_1^*(s_1,s_2))) = \bar{\pi}_2^*(s_1;4) = a_2$$
$$\sigma_2^*(s_2,s_2,s_1) = \bar{\pi}_2^*(s_1;r_0(s_2,\sigma_0^*(s_2))+r_1(s_2,\sigma_1^*(s_2,s_2))) = \bar{\pi}_2^*(s_1;5) = a_1$$
$$\sigma_2^*(s_1,s_1,s_2) = \bar{\pi}_2^*(s_2;r_0(s_1,\sigma_0^*(s_1))+r_1(s_1,\sigma_1^*(s_1,s_1))) = \bar{\pi}_2^*(s_2;3) = a_1$$
$$\sigma_2^*(s_2,s_1,s_2) = \bar{\pi}_2^*(s_2;r_0(s_2,\sigma_0^*(s_2))+r_1(s_1,\sigma_1^*(s_2,s_1))) = \bar{\pi}_2^*(s_2;4) = a_2$$
$$\sigma_2^*(s_1,s_2,s_2) = \bar{\pi}_2^*(s_2;r_0(s_1,\sigma_0^*(s_1))+r_1(s_2,\sigma_1^*(s_1,s_2))) = \bar{\pi}_2^*(s_2;4) = a_2$$
$$\sigma_2^*(s_2,s_2,s_2) = \bar{\pi}_2^*(s_2;r_0(s_2,\sigma_0^*(s_2))+r_1(s_2,\sigma_1^*(s_2,s_2))) = \bar{\pi}_2^*(s_2;5) = a_2 .$$

この問題は最適政策が状態空間 X 上でマルコフ性をもたない例となっている.

参考文献

[1] R. Bellman, *Dynamic Programming*, Princeton Univ. Press, Princeton, New Jersey, 1957.

[2] R. Bellman and E. Denman, Invariant Imbedding, *Lecture Notes in Operation Research and Mathematical Systems*, Vol.**52**, Springer-Verlag, Berlin, 1971.

[3] K. Boda and J.A. Filar, Time consistent dynamic risk measures, *Math. Meth. Oper. Res.* **63**(2006), 169-186.

[4] M. Bouakiz and Y. Kebir, Target-level criterion in Markov decision processes, *J. Optim. Theory Appl.* **86**(1995), 1-15.

[5] J.A. Filar, D. Krass, and K.W. Ross, Percentile performance criteria for limiting average Markov decision processes, IEEE Trans. Automat. Contr. **40**(1995), 2-10.

[6] T. Fujita and K. Tsurusaki, Stochastic optimization of multiplicative functions with negative value. *J. Oper. Res. Soc. Japan* **41**(1998), no. 3, 351–373.

[7] S. Iwamoto, Associative dynamic programs, *J. Math. Anal. Appl.*, **201**(1996), 195-211.

[8] 吉良知文・植野貴之・藤田敏治, 制御マルコフ連鎖における成長確率最大化について, 京大数理研講究録 **1682**(2010), 62-69.

[9] Y. Ohtsubo, Optimal threshold probability in undiscounted Markov decision processes with a target set, *Applied Mathematics and Computation* **149**(2004), 519-532.

[10] Y. Ohtsubo and K. Toyonaga, Optimal policy for minimizing risk models in Markov decision processes. *J. Math. Anal. Appl.* **271**(2002), 66-81.

[11] H. Simon, *Models of man*, New York: Wiley, 1957.

[12] M.J. Sobel, The variance of discounted Markov decision processes. *J. Appl. Prob.* **19**(1982), 794-802.

[13] 植野貴之・岩本誠一, 確率最適化における過去集積値と未来閾値について, 京大数理研講究録 **1207**(2001), 79-100.

[14] D.J. White, Mean, variance, and probabilistic criterion in finite Markov decision processes: A review, *J. Optim. Theory Appl.* **56**(1988), 1-29.

[15] D.J. White, Minimizing a threshold probability in discounted Markov decision processes, *J. Math. Anal. Appl.* **173**(1993), 634-646.

[16] C. Wu and Y. Lin, Minimizing risk models in Markov decision processed with policies depending on target values, *J. Math. Anal. Appl.*, **231**(1999), 47-67.